CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Repositori Institucional de la Universitat Jaume I

Feature selection in regression tasks using conditional mutual information ⋆

Pedro Latorre Carmona[1], José M. Sotoca[1], Filiberto Pla[1], Frederick K. H. Phoa[2], and José Bioucas Dias[3]

[1] Dept. Lenguajes y Sistemas Informáticos, Jaume I University, Spain
[2] Institute of Statistical Science. Academia Sinica. R.O.C.
[3] Instituto de Telecomunicações and Instituto Superior Técnico, Technical University of Lisbon
[latorre,sotoca,pla]@uji.es,fredphoa@stat.sinica.edu.tw,bioucas@lx.it.pt

Keywords: Feature Selection,Regression,Information measures,Conditional Density Estimation

Abstract. This paper presents a supervised feature selection method applied to regression problems. The selection method uses a *Dissimilarity matrix* originally developed for classification problems, whose applicability is extended here to regression and built using the conditional mutual information between features with respect to a continuous relevant variable that represents the *regression function*. Applying an agglomerative hierarchical clustering technique, the algorithm selects a subset of the original set of features. The proposed technique is compared with other three methods. Experiments on four data-sets of different nature are presented to show the *importance* of the features selected from the point of view of the regression estimation error (using Support Vector Regression) considering the Root Mean Squared Error ($RMSE$).

1 Introduction

Feature selection aims at reducing the dimensionality of data. It consists of selecting the most relevant features (attributes) among the set of original ones [12]. This step is crucial for the design of regression and classification systems. In this framework, the term relevant is related to the impact of the variables on the prediction error of the variable to be regressed (*target variable*).The relevant criterion can be based on the performance of a specific predictor (wrapper method), or on some general relevance measure of the features for the prediction (filter method). Wrapper methods may have two drawbacks [19]: (a) they can be computationally very intensive; (b) their results may vary according to initial conditions or other chosen parameters. On the other hand, filter methods allow sorting features independently of the regressor. Eventually, embedded methods try to include the feature selection as a part of the regression training process. In order to tackle the combinatorial search problem to find an optimal subset of features, the

⋆ This work was supported by the Spanish Ministry of Science and Innovation under the projects Consolider Ingenio $2010CSD2007-00018$, and EODIX $AYA2008-05965-C04-04/ESP$ and by *Fundació Caixa-Castelló* through the project $P11B2007-48$.

most popular variable selection methods seek to avoid having to perform an exhaustive search applying forward, backward or floating sequential schemes [9][16].

This paper presents a (filter type) feature clustering-based method aiming at finding a subset of features that minimizes the regression error. To do this, the conditional mutual information will be estimated to define a criterion of distance between features. This distance has already been used in [18] for feature selection in classification tasks. Thus, the main contribution of this paper is to establish a methodology to properly solve the estimation of this distance for regression problems where the relevant variable is continuous, through the assessment of the conditional mutual information.

The organization of the rest of this paper is as follows. Section 2 gives an overview of the theoretical foundations of this distance and solves its estimation in regression problems. Section 3 describes the experiments carried out, the databases used and the other feature selection methods in regression used in the comparison. Section 4 presents and discusses the regression results obtained. Finally, some concluding remarks are depicted in Section 5.

2 Feature selection for regression

2.1 The Minimal-Relevant-Redundancy criterion function

Sotoca *et al* show in [18] that if $\widetilde{\mathbf{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_m)$ is a subset of m random variables out of the original set of n random variables $\mathbf{X} = (X_1, \ldots, X_n)$, that is, $\widetilde{\mathbf{X}} \subset \mathbf{X}$, then, the decrease in mutual information about a relevant variable Y can be expressed as $I(\mathbf{X};Y) - I(\widetilde{\mathbf{X}};Y) = I(\mathbf{X};Y/\widetilde{\mathbf{X}})$. They also show that the decrease of mutual information of the original and the reduced sets with respect to the relevant variable Y is upper bounded by $I(\mathbf{X};Y) - I(\widetilde{\mathbf{X}};Y) \leq \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i;Y/\widetilde{X_j})$, where $I(X_i;Y/\widetilde{X}_j)$ is the conditional mutual information between the feature X_i and the cluster representative $\widetilde{X}_j$ and expresses how much information variable X_i can predict about the relevant variable Y that $\widetilde{X}_j$ cannot. This bound can be interpreted as a *Minimal Relevant Redundancy - mRR* criterion, meaning that the selected features will tend to be as independent as possible with respect to the information content of the relevant variable Y they are attempting to predict.

One way to find a solution to the minimization problem is to approximate this bound by a clustering process, where $\widetilde{\mathbf{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_m)$ are the representative features of each cluster. To solve this problem, given two variables X_i and X_j, the following function satisfies the properties of a metric distance:

$$D(X_i, X_j) = I(X_i;Y/X_j) + I(X_j;Y/X_i) = H(Y/X_i) + H(Y/X_j) - 2 \cdot H(Y/X_i, X_j) \tag{1}$$

The same metric will be considered here, using an agglomerative hierarchical approach based on a Ward's linkage method [20]. The number of groups is reduced at each iteration until m clusters are reached. For each resulting cluster, C_j, its representative feature $\widetilde{X}_j$ is chosen as the feature X_j with the highest value of the mutual information with respect to the continuous relevant variable Y, that is,

$$\widetilde{X}_j = \{X_j \in C_j \; ; \; I(X_j;Y) \geq I(X_i;Y); \quad \forall X_i \in C_j\} \tag{2}$$

2.2 Estimation of the Conditional Mutual Information for regression tasks

Given a set of N n-dimensional training samples $(\mathbf{x}_k, y_k)$, $k = 1, \ldots, N$ defined by the set of variables (features) $\mathbf{X} = (X_1, \ldots, X_n)$ where there is a dependency $y_k = f(\mathbf{x}_k)$, and for which a specific regressor can be applied, the conditional differential entropy $H(Y/\mathbf{X})$ can be written as [1]:

$$H(Y/\mathbf{X}) = -\int p(\mathbf{x}, y) \log p(y/\mathbf{x}) d\mathbf{x} dy \tag{3}$$

Considering that the joint probability distribution, $p(\mathbf{x}, y)$ can be approximated by the *empirical* distribution [15]: $p(\mathbf{x}, y) = \frac{1}{N} \cdot \sum_{k=1}^{N} \delta(\mathbf{x} - \mathbf{x}_k, y - y_k)$, where $\delta(\mathbf{x} - \mathbf{x}_k, y - y_k)$ is the Dirac delta function, and substituting $p(\mathbf{x}, y)$ into Eq. 3, we have:

$$H(Y/\mathbf{X}) = -\frac{1}{N} \cdot \sum_{k=1}^{N} \log p(y_k/\mathbf{x}_k) \tag{4}$$

Calculating Eq. 4 for one and for all pairs of two variables X_i, X_j, and substituting in Eq. 1, the *Dissimilarity matrix* of distances $D(X_i, X_j)$ can be obtained.

The assessment of $p(y/\mathbf{x})$ in Eq.4 is usually called Kernel Conditional Density Estimation (*KCDE*). This is a relatively recent active area of research that basically started with the works by Hyndman *et al* [8], among others. One way to obtain $p(y/\mathbf{x})$ is to use a (training) dataset $(\mathbf{x}_k, y_k)$ and a Nadaraya-Watson type kernel function estimator, as in [7], considering only the y_k training values that are paired with values $\mathbf{x}_k$:

$$\widehat{p}(y/\mathbf{x}) = \frac{\sum_k K_{h_1}(y - y_k) \cdot K_{h_2}(\|\mathbf{x} - \mathbf{x}_k\|)}{\sum_k K_{h_2}(\|\mathbf{x} - \mathbf{x}_k\|)} \tag{5}$$

where K_h is a compact symmetric probability distribution function, for instance, a gaussian kernel. In this case:

$$K_h(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} h^n |\Sigma|^{\frac{1}{2}}} \cdot \exp\left(-\frac{\mathbf{x}^T \Sigma^{-1} \mathbf{x}}{2h^2}\right) \tag{6}$$

where Σ is a covariance matrix of a n-dimensional vector, $\mathbf{x}$. There are two bandwidths, h_1 for the y kernel and h_2 for the $\mathbf{x}$ kernel. The accuracy in the estimation of the conditional density functions is dependent on the assessment of the (h_1, h_2) parameters. The most common way to establish this accuracy would be the *Mean Integrated Square Error* [17] which can be defined (in this case) as:

$$MISE(h_1, h_2) = \int \left[p(y/\mathbf{x}) - \widehat{p}(y/\mathbf{x})\right]^2 dy p(\mathbf{x}) d\mathbf{x} \tag{7}$$

However, the following cross-validated log-likelihood defined in [7] will be used here because of its lower computational requirements:

$$L(h_1, h_2) = \frac{1}{N} \sum_k \log(\widehat{p}^{(-k)}(y_k/\mathbf{x}_k) \cdot \widehat{p}^{(-k)}(\mathbf{x}_k)) \tag{8}$$

where $\widehat{p}^{(-k)}$ means $\widehat{p}$ evaluated with $(\mathbf{x}_k, y_k)$ left out. $\widehat{p}(\mathbf{x})$ is the standard kernel density estimate over $\mathbf{x}$ using the bandwidth h_2 in Eq. 5. It can be shown that maximizing the *KCDE* likelihood is equivalent to minimizing the *MISE* criterion. Substituting the Watson-Nadaraya type kernels into the $L(h_1, h_2)$ [7]:

$$\begin{aligned} L(h_1, h_2) &= \frac{1}{N} \sum_k \log \left[\left(\frac{\sum_{j \neq k} K_{h_1}(y_k - y_j) K_{h_2}(\|\mathbf{x}_k - \mathbf{x}_j\|)}{\sum_{j \neq k} K_{h_2}(\|\mathbf{x}_k - \mathbf{x}_j\|)} \right) \cdot \left(\sum_{j \neq k} \frac{K_{h_2}(\|\mathbf{x}_k - \mathbf{x}_j\|)}{N-1} \right) \right] \\ &= \frac{1}{N} \sum_k \log \left(\frac{\sum_{j \neq k} K_{h_1}(y_k - y_j) K_{h_2}(\|\mathbf{x}_k - \mathbf{x}_j\|)}{N-1} \right) \end{aligned} \tag{9}$$

3 Experimental validation

3.1 Methods

The proposed method, hereafter called MI_{Dist}, would consist of the steps described in Algorithm 1. Other three methods were used and compared to the method proposed here.

- Monteiro *et al* method [13] based on a Particle Swarm Optimization (*PSO*) strategy [10] (Particle-Swarms Feature Selection, $PS-FS$). It is a *Wrapper*-type method to make feature selection using an adaptation of an evolutionary computation technique developed by Kennedy and Eberhart [10]. For further details, see [13].
- Forward Stepwise Regression (FSR). Consider a linear regression model. The significance of each variable is determined from its t-statistics with the null hypothesis that the correlation between y and X_i is 0. The significance of factors is ranked using the p-values (of the t-statistics) and with this order a series of reduced linear models is built.
- Elastic Net (EN). It is a sparsity-based regularization scheme that simultaneously does regression and variable selection. It proposes the use of a penalty which is a weighted sum of the l_1-norm and the square of the l_2-norm of the coefficient vector formed by the weights of each variable. For further details, see [21].

3.2 Dataset description

Four datasets were used to test the feature selection methods, divided into three groups:

- Hyperspectral datasets. Two hyperspectral datasets corresponding to a *Remote Sensing* campaign ($SEN2FLEX2005$, [14]) were used.
  1. *CASI-THERM*. It consists of the reflectance values of image pixels that were taken by the Compact Airborne Spectrographic Imager ($CASI$) sensor [14]. Corresponding thermal measurements for those pixels were also performed. The $CASI$ sensor images are formed by 144 bands between 370 and 1049nm.

1. **Kernel width estimation**. Obtain, for each pair and tuple $(Y; X_i)$ and $(Y; X_i, \widetilde{X}_j)$, the pair of parameters (h_1, h_2) that maximize $L(h_1, h_2)$ (Eq. 9).
2. **Kernel Density Estimation**. Obtain the Watson-Nadaraya type Kernel Density estimators $K_{h_1}(y - y_k)$ and $K_{h_2}(\|\mathbf{x} - \mathbf{x}_k\|)$ applying Eq. 6
3. **Assessment of the A-Posterior Probabilities**. Estimate $\widehat{p}(y/\mathbf{x})$ as $\widehat{p}(y/\mathbf{x}) \leftarrow \frac{\sum_k K_{h_1}(y-y_k)\cdot K_{h_2}(\|\mathbf{x}-\mathbf{x}_k\|)}{\sum_k K_{h_2}(\|\mathbf{x}-\mathbf{x}_k\|)}$.
4. **Estimation of the Conditional Entropies**. Obtain, for each variable $X = X_i$ and every possible combination $X = (X_i, \widetilde{X}_j)$ the Conditional Entropies: $H(Y/X) \leftarrow -\frac{1}{N} \cdot \sum_{k=1}^{N} \log p(y_k/\mathbf{x}_k)$.
5. **Construction of the** *Dissimilarity Matrix* $D(X_i, \widetilde{X}_j)$. Obtain $D(X_i, X_j) = I(X_i; Y/X_j) + I(X_j; Y/X_i) = H(Y/X_i) + H(Y/X_j) - H(Y/X_i, X_j)$
6. **Clustering**. Application of a Hierarchical Clustering strategy based on Ward's linkage method to find clusters using $D(X_i, \widetilde{X}_j)$. The number of clusters is determined by the number of variables to be selected.
7. **Representative selection**. For each cluster the variable with the highest value of the mutual information with respect to the continuous relevant variable Y is selected.

Algorithm 1: Selection of variables using Mutual Information based measures

2. *CASI-AHS-CHLOR*. It consists of the reflectance values of image pixels that were taken by the $CASI$ and Airborne Hyper-spectral Scanner (AHS) [14] sensors. Corresponding chlorophyll measurements for those pixels were also performed. AHS images consist of 63 bands between 455 and 2492 nm.

- *Bank32NH*. It consists of 8192 cases, 4500 for training and 3692 for testing, with 32 continuous variables, corresponding to a simulation how bank-customers choose their banks. It can be found in the *DELVE* Data Repository.
- *Boston Housing*. Dataset created by D. Harrison et al [6]. It concerns the task of predicting housing values in areas of Boston. The whole dataset consists of 506 cases and 13 continuous variables. It can be found in the *UCI* Machine Learning Repository.

4 Results and discussion

One (h_1, h_2) kernel width pair was obtained for each $(y_k, \mathbf{x}_k)$ pair, applying an Active Set method [5] for non-linear multi-variable optimization with constraints. The starting values were fixed at: $h_{1,0} = h_{2,0} = \frac{1}{2\log(N)}$, as in [11], and the lower and upper bounds at $[h_{i,m}, h_{i,M}] = [0.1 \cdot h_{i,0}, 10 \cdot h_{i,0}]$, $i = 1, 2$. For the assessment of $p(y/\mathbf{x})$ when $\mathbf{x} = (X_i, X_j)$, the covariance matrix considered was diagonal: $\Sigma = diag(\sigma_i^2, \sigma_j^2)$, where σ_i^2 and σ_j^2 are the variance value of variables i and j, respectively, for the training set. For the selection of the variable that represents a cluster, Eq.2 was applied.

Support Vector Regression (SVR) [2] with a radial (gaussian) basis function was used for regression, and the Root Mean Squared Error ($RMSE$) as the performance criterion. For each one of the four datasets, an exhaustive grid search using equally spaced steps in the logarithmic space of the *tuning* SVR parameters (C, ε, σ) was made

Database	MI_{Dist}	PS−FS	FSR	EN	Friedman Test	Quade Test
CASI-AHS-CHLOR	**2.9160**	3.1266	4.5636	3.3524	6.53 (-)	7.24 (+)
CASI-THERM	**3.3267**	3.3892	3.6425	3.4381	0.08 (-)	0.02 (-)
Bank32NH	0.0961	0.0953	**0.0950**	**0.0950**	0.48 (-)	1.03 (-)
Boston Housing	4.4270	**4.3702**	4.8016	4.6252	1.43 (-)	1.44 (-)
Average	**2.6914**	2.7454	3.2757	2.8777		

Table 1. $RMSE$ Error over the first 5 variables. (+ = positive), (− = negative)

to select and fix the best parameters for each one of these datasets. A 10-fold cross-validation strategy was used to obtain the $RMSE$ error on the *Boston Housing* dataset.

Figure 1 shows the *RMSE* error given by the *SVR* method for the four datasets and the first 20 variables selected (13 variables for *Boston Housing*) by each one of the variable selection methods tested. Tables 1 to 3 show the $RMSE$ Error over the first 5, 10 and 15 variables (13 variables for *Boston Housing*) for all the methods and datasets selected. Friedman and Quade Tests [3], [4] were applied on the results with a confidence level of $p = 0.005$. The Fisher distribution critical value for the four methods and over the first 5, 10 and 15 variables (13 for *Boston Housing*) was set up, obtaining $F(3, 12) = 7.20, F(3, 27) = 5.36, F(3, 42) = 4.94$. The differences in $RMSE$ ranked for the four methods are not significant for the first 5 features, but they are for 10 to 15 features. Thus, the difference between the methods increases with the number of selected features, although in the case of *CASI-THERM* database the statistical tests are not significant.

The proposed method, MI_{Dist}, obtains better performance with respect to the rest of methods for all the cases (5, 10 and 15 variables) for the *CASI-AHS-CHLOR* and *CASI-THERM* datasets and for two out of the three (10 and 13 variables) for the *Boston Housing* dataset. *PS−FS* method is the second best one in most cases followed by the *EN* method, while *FSR* behaves worst. When averaging each method over the four databases, MI_{Dist} obtains lower values in the three tables.

The selection of the first one and the first two variables is better for MI_{Dist} compared to the rest of the methods. In this case, the clustering strategy plays an important role in the formation of different groupings of features, obtaining better results than a greedy selection algorithm as is the case of *FSR*. In the cases of *PS−FS* and *EN* methods the possible advantage of our method consists in a proper adjustment of the parameters from Nadaraya−Watson function estimator and its use through a distance metric in feature space that takes into account the internal relationships between features.

Database	MI_{Dist}	PS−FS	FSR	EN	Friedman Test	Quade Test
CASI-AHS-CHLOR	**2.3970**	2.5448	4.1506	2.8153	28.04 (+)	20.73 (+)
CASI-THERM	**3.1912**	3.2866	3.3585	3.2501	1.17 (-)	0.93 (-)
Bank32NH	0.0930	0.0913	**0.0910**	0.0913	4.14 (-)	6.21 (+)
Boston Housing	**4.2035**	4.3264	4.7022	4.8756	6.73 (+)	5.98 (+)
Average	**2.4712**	2.5623	3.0756	2.7581		

Table 2. $RMSE$ Error over the first 10 variables. (+ = positive), (− = negative)

Database	MI_{Dist}	PS−FS	FSR	EN	Friedman Test	Quade Test
CASI-AHS-CHLOR	**2.1964**	2.2922	3.7828	2.5497	49.13 (+)	30.43 (+)
CASI-THERM	**3.2055**	3.2771	3.3027	3.2308	2.17 (-)	1.68 (-)
Bank32NH	0.0920	0.0908	**0.0905**	0.0907	1.21 (-)	7.71 (+)
Boston Housing	**4.3171**	4.5162	4.7995	4.9491	7.84 (+)	8.08 (+)
Average	**2.4517**	2.5436	2.9950	2.7051		

Table 3. $RMSE$ Error over the first 15 variables (13 for *Boston Housing*). (+ = positive), (− = negative)

5 Conclusions

This paper presents a *filter*-type method to do feature selection for regression, using a *Dissimilarity matrix* based on conditional mutual information measures. This matrix is an extension for continuous variables of a Dissimilarity matrix used by Sotoca *et al* in [18] for classification. The method is compared against three other methods on four datasets of different nature, using the $RMSE$ Error given by the SVR technique. Authors are currently working on the performance analysis with other types of regressors and datasets and analyzing the effect of noise in the data on the performance of the selection strategy.

References

1. T. M. Cover, J.A. Thomas, "Elements of Information Theory", John Wiley & Sons Inc., 1991.
2. H. Drucker, C. Burges,L. Kaufman, L. Kaufman, A. Smola, V. Vapnik, "Support Vector Regression Machines", *in NIPS96*, pp. 155-161, 1996.
3. M. Friedmann, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance", *J. Am. Stat. Assoc.*, vol. 32, no. 200, pp. 675-701, 1937.
4. S. García, A. Fernández, J. Luengo, F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power", *Information Sciences*, vol. 180, pp. 2044-2064, 2010.
5. P. E. Gill, W. Murray, M. H. Wright, "Practical Optimization", Academic Press, 1981.
6. D. Harrison, D. L. Rubinfeld, "Hedonic prices and the demand for clean air", *J. Environ. Econ. Manag.*, vol.5, pp. 81-102, 1978.
7. M. P. Holmes, A. Gray, C. L. Isbell, "Fast kernel conditional density estimation: A dual-tree Monte Carlo approach", *Comput. Stat. Data Analysis*, vol. 54, no. 7, pp. 1707-1718, 2010.
8. R. J. Hyndman, D. M. Bashtannyk, G. K. Grunwald, "Estimating and visualizing conditional densities", *Journal of Computation and graphical Statistics*, vol. 5, no. 4, pp. 315-336, 1996.
9. A. K. Jain, R. P. W. Duin and J. Mao, "Statistical pattern recognition: A review", *IEEE Trans. PAMI*, vol. 22, no. 1, pp. 4-37, 2000.
10. J. Kennedy, R. Eberhart, "Particle Swarm Optimization", *in IEEE ICNN.*, pp. 1942-1948, 1995.
11. N. Kwak and Chong-Ho Choi, "Input feature selection by mutual information based on parzen window", *IEEE Trans. PAMI*, vol. 24, no. 12, pp. 1667-1671, 2002.
12. M. Dash and H. Liu, "Feature selection for classification", *Intelligent Data Analysis*, vol. 1, pp. 131-156, 1997.

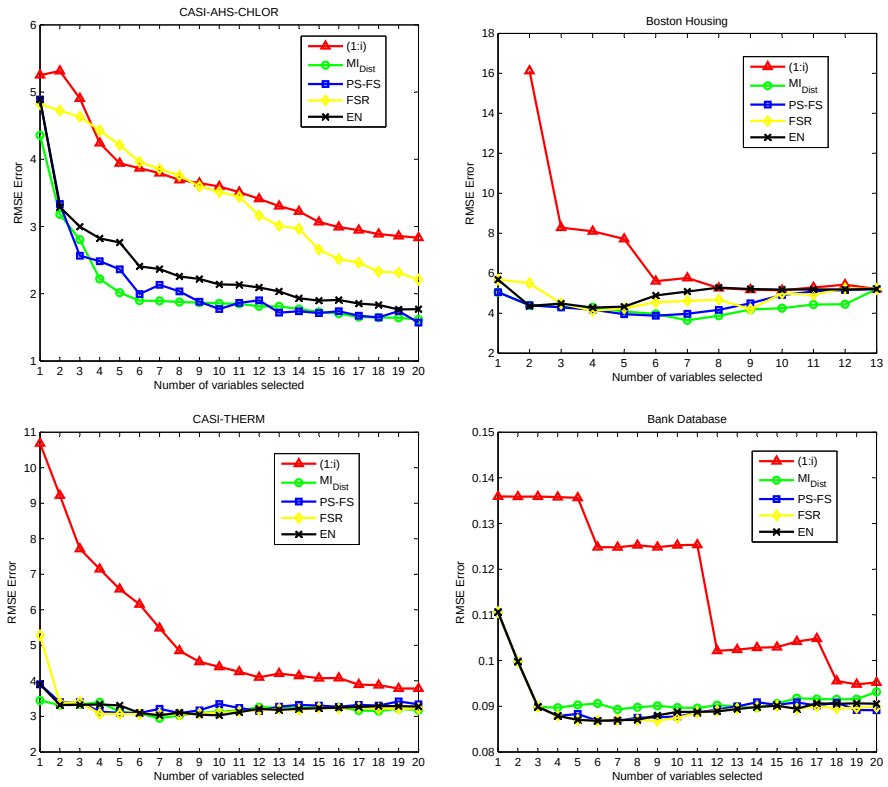

Fig. 1. $RMSE$ Error using SVR for the *CASI-AHS-CHLOR*, *Boston Housing*, *CASI-THERM* and *Bank*32*NH* datasets, respectively. The first point in the $(1:i)$ line for *Boston Housing* is not shown because it is of the order of ~ 2000

13. S. T. Monteiro, Y. Kosugi, "Particle Swarms for Feature Extraction of Hyperspectral Data", *IEICE Trans. Inf. and Syst.,* vol.E90D, no.7, pp. 1038-1046, 2007
14. J. F. Moreno, "Sen2flex data acquisition report", *Universidad de Valencia, Tech. Rep.*, 2005.
15. H. Ney, "On the relationship between classification error bounds and training criteria in statistical pattern recognition", *in IbPRIA03*, pp. 636-645, 2003.
16. P. Pudil, F. J. Ferri, J. Novovicova and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions", *Pattern Recognition*, vol. 2, pp. 279-283, 1994.
17. M. Rosenblatt, "Remarks on some nonparametric estimates of a density function", *Ann. Math. Statist.*, vol. 27, pp. 832-837, 1956.
18. J. M. Sotoca, F. Pla, "Supervised feature selection by clustering using conditional mutual information-based distances", *Pattern Recognition*, vol. 43, no. 6, pp. 2068-2081, 2010.
19. M. Verleysen, F. Rossi, D. Franois, "Advances in Feature Selection with Mutual Information", *Similarity-based Clustering*, M. Biehl et al. (Eds.), LNAI, no. 5400, pp. 52-69, 2009.
20. J. H. Ward, "Hierarchical grouping to optimize an objective function", *J. Am. Stat. Assoc.*, vol. 58, no. 301, pp. 236-244, 1963.
21. H. Zou, T. Hastie, "Regularization and variable selection via the elastic net", *J. R. Statist. Soc. B*, vol. 67, part 2, pp. 301-320, 2005.